科研智能化趋势下科研数据服务研究

张婧睿[1, 2]　　孙蒙鸽[1, 2]　　韩涛[1]
[1]（中国科学院文献情报中心 北京 100190）
[2]（中国科学院大学经济与管理学院图书情报与档案管理系　 北京 100190）

摘要：
[目的] 系统梳理和总结科研智能化趋势下科研数据在科研过程中的运行流程，挖掘其中潜在的科研数据需求，为新趋势下科研数据服务的转型发展提供思考。**[方法]** 在科研数据生命周期的理论指导下，以材料和化学领域为例分析科研数据在科研智能化研究中如何转变为知识的过程，构建了包括数据管理计划、数据产生与收集、数据处理与分析、数据生成与出版、数据存储与共享、数据再利用六大阶段的科研数据生命周期运行流程，挖掘科研数据的作用和潜在需求。**[结果]** 科研智能化研究表现出对多源异构数据集成、细粒度数据结构化、人机互动语言表示的探索、数据关联化挖掘和科研数据类型丰富化的需求特征。**[结论]** 建议未来科研数据服务发展加强高质量全面化领域数据网络建设、深化嵌入科研式数据服务、提升图书馆员领域知识和人工智能素养、重视文本型数据中实验信息的挖掘、关注人机互动语言的探索。
关键词：科研智能化　 科研数据　 科研数据服务

Research on Scientific Research Data Services under the Trend of Intelligent Scientific Research

Zhang Jingrui[1, 2]　 Sun Mengge[1, 2]　 Han Tao[1]
[1](National Science Library, Chinese Academy of Science, Beijing 100190, China)
[2] (Department of Library, Information and archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract:
[Objective] Systematically sort out and summarize the operation process of scientific research data in the scientific research process under the trend of Intelligent Scientific Research, mine the potential scientific research data demand, and provide thinking for the transformation and development of scientific research data services under the new trend.

[Methods] Under the guidance of the theory of scientific research data life cycle, taking the field of materials and chemistry as an example, this paper analyzes how scientific research data can be transformed into knowledge in the intelligent research of scientific research, and constructs six stages of scientific research data life cycle operation process, including data management plan, data generation and collection, data processing and analysis, data generation and publication, data storage and sharing, and data reuse, so as to explore the role and potential needs of scientific research data.

[Results] The research on intelligent scientific research demonstrates the exploration of multi-source heterogeneous data integration, fine-grained data structuring, human-machine interaction language representation, data association mining, and the enrichment of scientific research data types.
[Conclusions] It is recommended to strengthen the construction of high-quality and comprehensive data networks in the field of scientific research data services in the future, deepen embedded scientific research data services, enhance the knowledge and artificial intelligence literacy of librarians in the field, attach importance to the mining of experimental information in textual data, and pay attention to the exploration of human-machine interaction language.
Keywords: Intelligent scientific research　Research Data　Research Data Services

1　引言

近年来，自成熟的人工智能(AI)技术不断应用到具有挑战性的基础科学研究后，极大提升了科研效率，引发了一场改变科学研究态势的热潮，科学研究的知识发现主体也由科研人员转变为智能科学家，研究对象由传统的实验对象转变为科研数据。

科研数据是指各学科领域在科研活动的全过程中产生的各类数据[1]，包括以文本形式呈现的成果数据--科技文献、专利和基础研究、应用研究、试验开发、观测检验等产生的科学数据。目前，科研数据是以人类可理解可学习形式和数字化存储形式呈现，人工智能技术无法准确提取数据中隐含的科学规律，只有受到特定领域约束的科研数据才能供给人工智能学习。

数据密集型科研范式下科研数据存在增长速度快、规模巨大、来源和格式多样化的特点，为数据管理保存、集成关联、共享利用等带来了挑战。因此，在发展过程中着重强调了科研范式的数字化转型，提出了构建开放共享科学数据库或平台以及建立数据之间显性关联的需求，以实现科研数据的发现、访问、集成和分析，建立原本不相关领域之间的数据关系，促进知识发现[2]。由此，世界各国持续发力，以科研数据集成共享为出发点布局相关的战略，例如欧盟在第七框架计划(7th Framework Programme, FP7)(2007-2013年）启动全球科学数据基础设施建设项目 GRDI 2020(Global Research Data Infrastructures)，将科研数据集成共享纳入科研计划中[3]。

随着海量科研数据的积累和 AI 技术在科学研究中的深度融合催生了科研智能化研究范式。新范式下科研数据呈现出多源异构、多维度、关联性的复杂性特点，数据密集型科研范式下共享集成的科研数据具有分散性、领域数量差异大、质量参差不齐、标准和格式不一致等问题，为 AI 模型的可读、可利用和可理解带来挑战。基于此需求美国开始部署相关政策机制以把握新趋势下的科研数据主导权，例如 2020 年 5 月美国国立卫生研究院（NIH）共同基金启动“Bridge2AI”计划以生成机器可理解的统一标准化的生物医学和行为数据集和开发自动化工具加速标准化数据集生成为目标；2021 年美国材料基因组计划面向科研智能化趋势确立了统一规范化材料数据基础设施、推动材料数据开放共享和元数据标准统一的战略目标以充分发挥材料数据

在人工智能研发领域的力量[4]。

面向科研范式的变革，国家科研数据政策和计划的部署为科研智能化研究的发展和进步提供了数据服务保障。在科研智能化研究实践中发现新趋势下科研数据服务由领域科研人员、数据科学家以及信息服务人员组成，其中领域科研人员和数据科学家从科研智能化研究需求和前沿人工智能技术出发不断探索和参与智能化研究趋势下的科研数据研究工作，构建高质量、细粒度、多模态的科研数据库以满足 AI 模型的数据需求；与其相比，信息服务人员在科研智能化研究趋势下的科研数据服务仍然集中在传统的文献元数据层面的组织、前沿技术的咨询和培训服务方面，虽然初步探索了细粒度数据的抽取和知识服务转型的研究，但与领域科研人员和数据科学家相比在新趋势下科研数据服务的探索度和参与度存在劣势，未在已有工作中凸显文献信息服务机构的数据和技术结合的优势。

综上可知，科研智能化研究已经认识到科研数据的重要性，并开展了相关的政策和计划部署。科研数据在科学研究中是一个生命周期运动的过程，现有研究缺乏对新趋势下科研数据生命周期运行流程、作用和需求的探索，以了解新趋势下科研数据发展的现状。因此，本文从科研智能化实践研究工作出发，基于科研数据生命周期视角总结梳理了新趋势下科研数据在科学研究中的生命周期运行过程、作用和潜在需求，并据此为新趋势下文献信息服务机构科研数据服务的发展提供思路。

2　科研智能化趋势下科研数据生命周期运行过程分析

科研数据在科学研究中的运行流程是一个目标驱动的 DIKW 模型，也即是在研究目标的驱动下科研数据向信息、知识、智慧的转变过程，以支持研究目标的实现。本章节以科研智能化研究前沿领域—材料和化学领域为例，探究科研智能化趋势下科研数据生命周期运行过程，主要是分析数据如何转变为最终的智慧（图 1），该流程框架的构建是从研究实践角度出发，基于数据生命周期理论对各环节和流程进行详细阐述和分析。

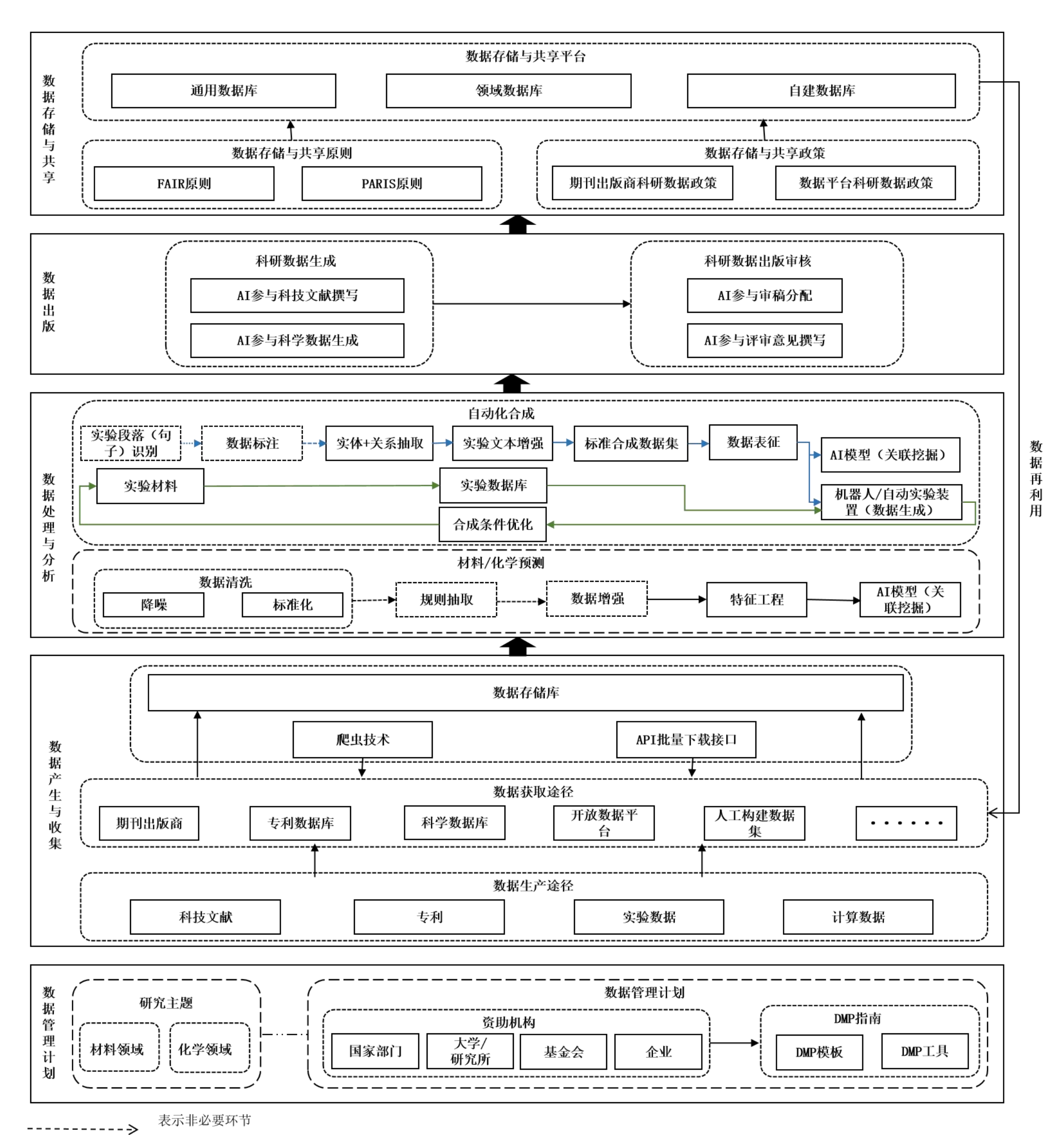


图 1 科研智能化趋势下科研数据生命周期运行过程

2.1 数据管理计划（DMP）

研究主题的确定为科研项目工作树立了一座“导航塔”。在确定研究主题之后，需要制定辅助科研项目数据管理工作的计划。数据管理计划（DMP）是一个在整个研究项目生命周期内以描述项目数据收集、记录、管理和发布的动态文档，包括创建、记录、访问、存储和共享的技术、方法和政策[5]，在科研数据和成果的管理和审查中发挥着关键作用。开放科学大背景下，科研资助机构作为推动开放获取的主力军，发布了科研项目申请时提交数据管理计划的政策，促进科研数据开放共享[6]。

对比国内外资助机构 DMP 基础上发现，国外资助机构 DMP 服务提供相应的工具和模板，对资助项目的 DMP 所包含内容进行具体拆解和说明，重视科研数据存储后的共享与访问、安全与伦理及 DMP 成本管理。国内 DMP 以国家科技部和中科院为主体对资助计划或项目所产生的数据进行管理，未提供相应的 DMP 工具与模板，重点在于科研数据的汇交与管理，明确了数据相关主

体职责及汇交途径，缺乏对后续共享与访问、成本管理的重视。

2.2 数据产生与收集

数据产生与收集阶段是支撑科学研究的基础环节，科研人员基于研究目标和数据可用性选择合适的数据类型和数据获取途径以获取所需研究数据内容，后续科学研究提供“燃料”（图 2）。

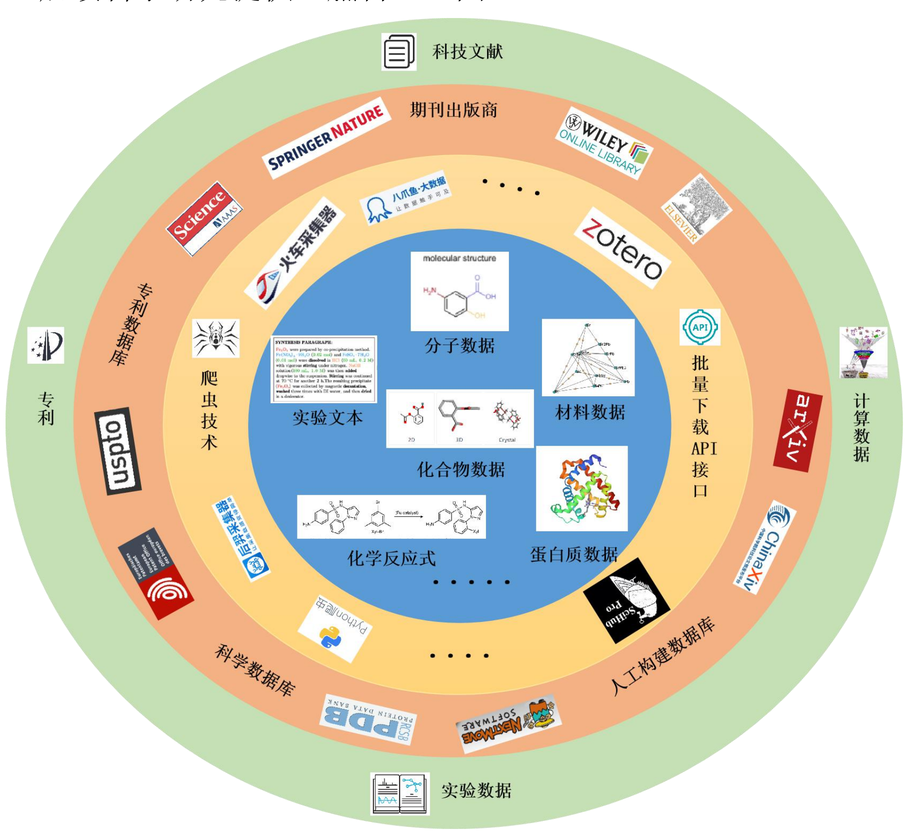


图 2 科学研究中数据产生类型与获取途径

（1）数据生产途径

数据生产是开展科学研究的重要起始阶段，为科研智能化研究积累了大量可用数据。现有数据生产途径主要有科技文献、专利、实验数据和计算数据四大类，科技文献和专利文本数据通过人工抽取、半自动化和自动化方法抽取其中的元数据、实体及属性值、表格、图像和实验段落数据；实验数据是科研人员在观察、实验、调查过程中收集和生成的数据或借助电子记录实验本（ELN）工具数字化记录相关数据；计算数据是科研人员借助高通量筛选、计算平台或 AI 模型工具生成的模拟或预测数据。

（2）数据获取途径

数据获取途径是科研人员根据研究主题、数据质量、数据结构、数据易获取性的特点选择适合的数据集构建途径。现有数据获取途径主要包括期刊出版商、专利数据库、科学数据库、开放数据平台和人工构建数据集五大途径。

从数据来源和开放性出发，商业性期刊出版商由于文献数据收集范围广、结构化质量高成为科研人员的首选，常用数据库有 Wiley、Elsevier、Scoups 等，材料和化学领域常用特色数据有 ACS、the American Chemical Society、the Royal Society of Chemistry 等；开放数据平台因其数据易

获取性和数据积累量大的优势支持科研智能化研究，典型数据库包括 Arxiv 文献数据平台和 Figshare、Zenodo 等通用性数据共享平台。

总体而言，科研智能化研究是以文本数据、数值数据和图像数据为核心，从商业性和开源性平台获取所需数据。智能化研究对高质量数据的要求使得科研人员将数据收集途径转向期刊出版商，但存在数据库使用费用较高的缺点，对小型科研机构或团队获取数据不友好；开源性收集途径由于其数据易获取性为多数研究者青睐，但存在数据结构不一致、数据质量低的问题。

（3）数据收集方法与数据自生成式集成

数据收集方法是在科研人员确定数据获取途径后借助一定的技术批量收集所需数据。科研智能化研究所需数据量大，科研人员需要借助采集软件、爬虫代码或数据平台提供的 API 接口批量爬取或下载所需数据，以提高数据收集效率。

采集软件是已经封装好的爬虫平台，该软件属于“傻瓜式”采集模式，可用于网络数据采集和文献数据采集，例如八爪鱼采集器、Zetero、SciHub Pro 等；现有爬虫代码是基于 python 框架编写的，典型工具包有 Scrapy、BeautifulSoup、Requests-HTML、Selenium 等；数据自生成式集成是在机器人化学家或材料学家做完实验后将实验材料的组成、比例和实验结果数据存储到数据库或表格中以供 AI 模型学习使用。

2.3 数据处理与分析

数据处理与分析是科研智能化研究的核心环节，目的是保证输入数据质量和结构的一致性，转化为机器可理解形式，以建立研究目标与数据之间的关联。从使用的数据类型和研究目的出发将材料和化学领域的数据处理与分析分为以文本型数据为核心、以实验材料为核心和以数值型数据为核心的三大类数据处理与分析模式。

（1）以文本型数据为核心的数据处理与分析模式

以文本型数据为核心的数据处理与分析模式是从科技文献和专利文本中抽取实验关键元素、实验条件进行组合并将其转化为机器可读形式，主要包括实验段落（句子）识别、数据标注、实体和关系抽取、实验文本增强、数据表征和关联挖掘/数据生成六大处理环节。

在科技文献和专利中存在大量与实验合成不相关的冗余文本信息，增加了实验合成信息提取的难度。因此，需要确定与实验合成相关度高的段落（句子），缩小文本提取的空间。识别方法包括基于规则的方法和基于机器学习/深度学习的方法两类。基于规则的方法需要研究人员事先了解实验段落（句子）的关键特征以构建识别规则，例如特定领域的实验合成物质、属性标识符、指定值等，简单规则的构建以关键词匹配[7]为代表，复杂规则的构建以模式匹配[8]和正则表达式[9]为代表。该方法易于理解和解释，研究人员可以快速实验并修改，但当实验段落存在大量变量或约束条件变多时其灵活识别能力差。与基于规则的方法相比，基于机器学习/深度学习的方法具备较强的自主学习和灵活适应能力，通过学习实验段落（句子）特征进行分类，可分为基于传统机器学习的方法和基于深度学习的方法。基于传统机器学习的实验段落（句子）方法以分类方法为核心，需要少量人工特征标注数据，包

括适用于高维度二分类方法的逻辑回归分类[10]、适用于低维度多分类的随机森林方法[11]和适用于高维度多分类的贝叶斯方法[12]。基于深度学习的方法以其学习的速度和精准度突出，需要大量训练数据自主学习实验段落特征，但其调参工作复杂，模型可解释性差，包括适用于捕获长序列语义关系的 RNN 模型[13]和适用于自监督并行计算的 BERT 模型[14]。

标注数据是高性能模型学习的基础，帮助模型理解上下文信息，文本型数据抽取训练标签数据稀缺，在对数据进行抽取之前需要对实验段落进行标注以构建抽取模型所需的训练数据。随着 AI 模型发展对数据标注的重视，专业化的数据标注团队和平台相继出现，现有实验文本标注以众包标注和人工标注形式为主，知名众包标注平台以国外 Amazon Mechanical Turk（MTurk）为主，避免出现团队标注效率低的问题，人工标注通过领域专家对小型数据集标注，适用于标注数据量小的情况。

实体和关系抽取是从实验段落中识别产物、反应物、溶剂等实体和相应的实验条件及实体-实体和实体-实验条件之间的关系。实体和关系抽取方法包括集成抽取工具和基于深度学习的方法两类。集成抽取工具有 OSCAR4[15]、ChemicalTagger[16]、ChemDataExtractor[17]，其中 OSCAR4 适用于化学实体识别工作，以纯文本作为输入利用集成的正则表达式、词典和最大熵马尔可夫模型 （MEMM）三类实体识别器识别化学实体；ChemicalTagger 专注于专利的实验部分，以文本字符串作为输入利用集成的 OCSAR 工具和基于正则表达式的规则方法识别化学实体，并结合基于语法结构的短语解析器来识别操作短语和实体之间的关系以生成结构化的反应路径图，该工具依赖于人工构建的规则，对语言使用或预处理引入噪声敏感，在科技文献等非专利数据上的可扩展性差；ChemDataExtractor 是一个端到端的文本挖掘工具，对 PDF、HTML 和 XML 输入文件利用集成的条件随机场（CRF）、基于规则的短语解析器和表解析器从科技文献文本和表格中提取化学实体、属性、测量值和程序以构建数据集。基于深度学习的方法包括基于序列的抽取方法--Bi-LSTM-CRF[18]和基于预训练语言模型的抽取方法--BERT-CRF[19]。基于序列的抽取方法--Bi-LSTM-CRF 利用 Bi-LSTM 捕获长文本句子中单词的上下文语义关系，结合 CRF 模型预测输入句子的最佳标签链。基于预训练语言模型的抽取方法--BERT-CRF 利用语言模型有效获取文本中的上下文信息并通过抽取任务进行监督微调，结合 CRF 模型进行抽取。

在缺乏训练数据的情况下需要扩充已有数据量添加负样本数据满足模型学习需求，以提高模型的泛化能力，避免出现欠拟合或过拟合现象。文本型数据的增强是对实验合成序列的“改造”，通过替换序列中化合物名称、数量、时间、温度、体积等实体和属性以增强实验合成数据。

数据表征是数据和算法模型间的连接点，将实验合成数据转为机器可理解形式。数据表征方法主要包括 Word2Vec、EMLo、BERT 三类，其中 Word2Vec 由于其语义向量表征低维度的特点在科研智能化研究中常使用，但该模型生成的词向量属于静态表征，不能解决同义词问题；EMLo 模型采用双层双向的 LSTM 捕获上下文信息进行编码，属于动态表征，但由于 LSTM 本身的长距离依赖性问题无法捕获长序列，并且该模型不具备并行处理的能力；BERT 模型

是基于深度双向Transformer的预训练模型，通过利用单词上下文信息表征，属于动态表征解决了一词多义问题，该模型具备强大的并行运算和迁移学习能力常用于领域模型的预训练。

最后根据构建的标准实验数据集利用机器人自动合成新数据或利用AI算法学习实验物质及属性、实验条件参数之间的关联为合成预测服务。数据关联方法主要以属性预测为核心，主要包括分类、回归传统机器学习方法和数据生成的深度学习方法，传统分类和回归机器学习方法有适用于离散属性预测的支持向量机（SVM）[20]和随机森林回归模型[21]以及适用于连续属性预测的高斯回归过程[9]；深度学习方法适用于实验合成条件和目标联合概率分布学习的变分自编码器[13]。传统机器学习方法适用于人工提取特征充分、数据和计算资源受限的情况，模型方法直观、易于实现，例如科研人员根据一组给定的沸石合成参数数据（包括组合元素相关的数值类型—数值、范围或变量；合成操作动作和条件）利用随机森林模型对沸石材料结构特性进行建模[21]；深度学习方法直接对原始高维数据进行隐含结构与关联性的挖掘突破了人工提取特征的局限性，但存在模型消耗资源多、结构复杂和不可解释性，例如科研人员借助变分自编码器将合成参数压缩为低维表示建立合成条件和前体材料之间的概率分布关系[13]。

（2）以实验材料为核心的数据处理与分析模式

以实验材料为核心的数据处理与分析模式是在科研人员提供的实验材料基础上利用机器人/自动实验装置对实验材料进行组合实验构建实验结果数据空间，由AI模型构建实验组合与结果之间的函数关联以学习并迭代优化实验组合条件，该模式的核心环节在于合成条件优化学习，以寻找最佳实验组合条件。

从数据空间构建是否基于已有知识的角度可分为自生成式自动合成和学习式自动合成。自生成式自动合成是指机器人根据已有材料进行结合合成后产生结果建立了合成材料、条件和实验结果之间的函数关系，以供给模型进行迭代优化，其迭代优化数据是以实验材料的组合和机器实验结果为核心。Burger, B等[22]研制了一个可移动AI化学机器人，利用16个化学样品进行实验并使用气相色谱仪分析实验结果，基于实验样品和结果数据利用贝叶斯优化算法进行迭代学习；学习式合成是指根据已有实验合成数据空间筛选合适的实验材料组合条件指导机器人开展自动合成。Coley C W等[23]基于反应转化规则空间，利用神经网络模型筛选可用目标分子结构并评估反应质量。实验材料组合式的处理与分析以反应条件优化为核心环节，包括适用于二分类的线性判别分析（LDA）[24]和适用于属性独立多分类的贝叶斯优化算法，但线性判别分析存在不适用于类别不均的数据分析问题，贝叶斯优化算法也存在计算量大，调参复杂的问题。

（3）以数值型数据为核心的数据处理与分析模式

以数值型数据为核心的数据处理与分析模式通过利用AI模型的数据表征学习和计算能力探索材料/化学数据结构、组成和特性之间的复杂空间关联，以对实现对三者的预测或生成，主要包括数据清洗、规则抽取、数据增强、特征工程和关联挖掘五大环节。

数据清洗是剔除数据集中存在缺失、错误和重复数据或对数据不同表示形式进行统一化。数据清洗方法包括降噪和标准化两类，降噪基于统计学方法，利用神经网络灾难遗忘策略[25]（catastrophic forgetting）剔除学习率低的异常化学反应数据，提高了前向预测和逆向合成模型的性能。此外，由于不同数据库选择描述分子结构的原子起点不同导致不同分子结构 SMILES 表示的产生，因此需要将其转化为统一的规范化格式以提出重复分子，常用工具包包括 Python 工具包--RDKit 和 Java 工具包--CDK[26]。RDKit 工具包还可以在分子标准化基础上进一步计算分子描述符的功能，例如化合物结构相似性计算、分子构象优化、分子指纹生成等。CDK 工具包在数据规范化基础上还可以搜索化合物子结构、3D 图像生成、分子指纹生成等。

规则抽取是在学习化学反应中反应物和产物之间的原子映射信息基础上识别潜在的反应中心。现有抽取技术是基于深度神经网络的无监督方法，典型模型是 Transformer 模型，由于其不依赖于标注数据的无监督特性和对不平衡反应类型的适应性在规则提取中表现出巨大的潜力，例如 Transformer 模型从未标注化学反应数据中学习了原子在化学反应中的排列变化规律以提取反应规则[27]。

数据增强借助采样方法扩充小样本材料/化学数据。数据增强方法包括基于神经网络的数据增强方法、主动学习和迁移学习方法[28]。基于神经网络的方法通过无监督学习采样生成大量新数据，包括生成对抗网络[29]、变分自动编码器[30]。主动学习[31]利用机器学习从大量未标注数据选取有价值样本进行采样以代表大量未标注数据。迁移学习通过迁移相关领域的知识提高了模型对小数据的预测性能，如 Gupta 等[32]基于 ElemNet 模型在 OQMD 源数据集进行预训练，最后迁移至目标 JARVIS 数据集中进行微调材料属性。

特征工程是选择与研究目标相关的数据描述符表征材料/化学数据特征供给 AI 模型学习。特征工程包括特征选择和特征转换两大类，特征选择是指从高维度材料任务相关特征中去除冗余特征降低特征空间维度，以提高模型的预测精度和泛化能力，包括过滤式、包裹式和嵌入式三类。过滤式方法是基于统计学和互信息的方法对特征的重要性进行等级排名，该方法计算时间效率高，但未考虑特征之间的相关性，常用方法包括相关系数[33]和互信息[34]。包裹式方法在特征选择过程中结合了监督学习算法对特征子集进行评估，在评估过程中考虑特征之间的相关性和依赖性，但存在高维特征空间计算复杂度高的问题[35]，以支持向量机（SVM）-递归特征消除（RFE）方法[36]为代表。嵌入式方法嵌入到机器学习模型中，特征选择和模型训练过程无明显区分，常用方法包括基于惩罚项的方法和基于树的方法[37]；特征转换是指将高维特征空间映射到低维特征空间，实现特征降维，包括主成分分析（PCA）、线性判别分析等方法[38]。

关联挖掘通过探索结构、组成和性能之间的关联以满足预测需求。关联关系挖掘的方法包括以卷积神经网络构建的图结构关联模型和以 Bi-LSTM 和 Transformer 模型为基础构建的 Seq2Seq 深度学习模型，后者凭借其长序列远距离依赖的学习优势构建了化学领域反应物、试剂、催化剂和生成物的“翻译”关系。卷积神经网络擅长处理图像数据，能够对化合物结构图进行关联，

将原子和化学键表示为分子图中的节点和边，识别反应物和产物原子对之间的化学键变化，以建立反应物和产物之间的关联[39]；seq2seq 模型擅长处理文本类“翻译”问题，其并行计算优势提高了模型的效率，该模型将非数值类型的化学式转化为机器可识别形式，例如通过物理化学描述符、分子指纹等方式将分子表示为字符串，建立产物字符串和反应物字符串之间的关联以实现化学反应的逆向合成路线预测目标[40]。

2.4 数据生成与出版

数据生成与出版是指成果数据和科学数据的生成与出版工作。

（1）AI 参与科研数据生成工作

科研智能化趋势下 AI 模型也积极参与到研究论文的撰写工作中，主要涉及论文标题、摘要和论文生成任务。以最新发布的 ChatGPT[41]为典型代表，利用其生成式 AI 模型的优势基于庞大的领域文本训练数据集，能够根据输入主题和关键词参与完整论文生成过程，包括论文写作角度和思路、研究方法或工具查询、论文大纲、相关参考文献资源、生成完整论文内容、润色完善论文内容、期刊遴选等。

科学数据的生成不拘泥于研究人员提交的实验数据和高通量计算数据，也涉及科技文献和专利中包含的相关数据。早期科学数据库对科技文献和专利以手工摘录方式为主，随着海量文献和专利的发布和积累，手工方式显露出耗时且成本高的缺点，数据库构建也开始转向自动化方式。计算数据不同于以往的高通量筛选和计算产生的模拟数据，科研智能化趋势下计算数据是以 AI 模型计算产生的大量数据。此外，AI 模型也参与到代码生成工作中借助生成式 AI 模型进行预训练和微调以实现生成任务。

（2）AI 参与科研数据出版审核工作

面对指数级增长的科学出版物投稿量，对出版工作是一个巨大的挑战，其高质量评审工作是一个耗时耗力的过程，为解决出版繁重的评审压力并提高科研数据出版速度和效率，引入了机器学习模型参与审稿任务分配和评审意见撰写工作。如 Charlin 等[42]设计了投递论文分配工具—Toronto Paper Matching System（TPMS），通过比较投稿论文和审稿人已发表研究成果（代表审稿人的专业知识）之间的文本以计算投稿论文和审稿人专业知识之间的相关性。Yuan 等[43]利用 BART 预训练模型学习国际表征学习大会（ICLR）和 NeurIPS 会议论文与其评审意见之间的“评审翻译规律”。

2.5 数据存储与共享

数据存储是对科研项目完成后产出的科研成果与相关科研数据进行有序化管理，以实现科研数据的可发现、可获取和可重用。数据存储与共享政策主要是由资助机构和期刊出版商直接规定，其中资助机构是对资助项目衍生的研究论文和相关科研数据汇交和共享进行规定，以国家层面的政策为代表，包括欧盟 Horizon 2020 政策、中国的《科技计划形成的科学数据汇交技术与管理规范》等。期刊出版商是在获取研究论文的转让权后针对论文相关的科研数据发布了存储与共享管理政策。两者的最终目标是实现适应科研智能化研究的科研数据存储与共享平台的建设。

在数据存储与共享原则方面，现有科研数据存储与共享平台是基于 FAIR

原则侧重于数据发布与共享，在利用方面稍显不足，以数据汇聚/汇交为主要数据共享模式，不适用于科研智能化研究对多源异构和关联性数据存储与共享需求。面对数据融合和关联需求，PARIS 共享利用原则应运而生，从机器可处理分析、在线问答访问、数据安全可靠、数据关联与迁移性以及数据的有效供给五大方面出发，解决了科研数据分布式、孤岛化、差异化等问题以实现科研数据的高质量供给需求[44]。

在数据存储与共享政策方面，开放数据平台发布的科研数据政策相较于期刊出版商存在规范性差、限制性小的问题，对数据共享无强制性要求，并且没有制定共享数据的相关规范。与国外期刊出版商相比，我国期刊出版商（以中国科学出版社为代表）的科研数据政策规定较为泛化，尤其在数据存储库指南方面规定是借鉴国外相关出版商发布的相关政策。

在数据存储与共享平台建设方面，包括通用性、领域性和自建数据库三类，其中通用性和领域性数据库以知名开放性数据平台为主，如 Figshare、GitHub、PubChem 等。此外，科研智能化研究存在现有科研数据及其结构与研究需求不匹配的情况以及对研究预测结果的批量数据存储和共享需求，因此，科研人员通过自建数据库来促进科研数据存储和共享，典型案例是中科大化学机器人研究工作[45]。中科大研究团队为满足化学机器人对文献中化学反应知识的学习，构建了存储有 1120 万个包括反应物和生成物结构、名称、试剂、溶剂、催化剂、反应温度等环境参数的化学反应数据库。

2.6 数据再利用

数据再利用是对研究成果的二次利用，支持新一轮的数据分析与挖掘工作，是新一轮智能化研究的起点。从研究成果再利用角度出发，科研智能化研究本身就是对已有数据的再开发和使用，文本文献的再利用表现为新一轮科技文献的深度挖掘和专利文本挖掘数据的再利用，科学数据的再利用表现为数据库与已发布数据集的重复利用。从数据集重复使用角度，科研智能化研究中科技文献数据集由于其易获取性和内容丰富性已经成为数据再利用的核心。科学数据的再利用通过借助数据共享平台对数据库数据复用率进行分析，也即是通过数据的浏览、下载和引用量评估数据集的价值性和新颖性，例如 Figshare、Zenodo 等在数据集界面提供数据统计服务。

3 科研智能化趋势下科研数据作用

AI for science 趋势是在数据密集型科研范式下萌生并发展，数据是研究的基础和发现的源泉，AI 技术是研究的发动机。科研智能化以 AI 模型为技术核心，模型参数需要训练数据以捕获多样化领域知识特征，即挖掘已有知识空间的关键信息，构建知识路径。数据量越大，AI 模型学习到的隐含关键信息越全面，关联规则也越准确。因此，高质量、正负样本结合、多源异构性、结构化科研数据对推动科研智能化研究发展具有重要作用。

（1）高质量的科研数据是科研智能化研究准确率的“加速器”

人工智能和机器学习领域权威学者多次强调：“以数据为中心的 AI”，智能科学家也同样重视科研数据质量对知识发现的影响，持续探索构建高质量数据集的方式和技术。高质量数据不仅能够降低数据采集和预训练环节的

复杂度，而且也能够提高 AI 模型的性能。

现有知名科研数据库存在多样化的收集途径，对科研数据的质量把控依旧存在不足之处。比如，科研数据的收集未充分考虑人工智能需求，存在数据冗余、标注数据较少、数据一致性或标准化等问题，不适合 AI 模型学习。因此，科研智能化研究仍然需要关注高质量领域数据的处理和构建。

（2）正样本指示模型学习的方向，负样本设定模型学习的范围

科研数据中的正样本数据的作用在于训练 AI 模型学习数据中存在的共性特征；负样本是指科学实验中的非成功数据或低质量数据，也称为阴性数据，起到对比区分的作用，划定共性特征学习的边界，避免模型重复性捕获错误特征，以改进 AI 模型知识发现中存在的错误。

现有数据库中的科研数据大多为正样本数据，不能满足科研智能化研究对负样本数据的需求，负样本挖掘仍然依靠科研小组收集自身研究团队在科研过程中的“失败数据”，负样本数据收集耗时、数据量较小。因此，科研智能化研究需要重视负样本数据的收集、存储和发布。

（3）多源异构数据是科研智能化研究全面性的“保护舱”

科研智能化趋势下的知识发现过程是一个复杂问题求解过程，需进一步分解为不同层级子问题简化求解复杂度。在问题目标明确基础上，分解问题求解过程涉及数据层级及属性参数，来构建求解函数。异构数据拓展了对问题理解的角度或层级，多源数据丰富了不同层级属性参数信息，例如材料本身的性质特征与原子尺度的原子结构、电子结构、离子输运垒等数据相关，材料性质对外界环境的相应与外场条件的变化存在函数关系等[46]。

多源异构性是指科研数据的来源的多样性与格式、呈现形式的异构性。首先，科学数据主要是从科技文献和专利文本的实验文本、表格、图像中抽取，文本、表格和图像表现出不同的结构特征；其次，从文本中抽取的科学数据不仅包括数值型数据还包括图像型数据、三维立体结构数据等。这些多源多尺度数据能够助力智能科学家多渠道多途径了解信息并挖掘数据关联，以材料领域为例，多源多尺度数据助力探索微观-介观-宏观尺度的表征与关联关系。

（4）数据结构化是实现人机互动的桥梁

数据结构化是指抽取后的数据需按照一定的层次和语义结构组合，形成易于理解和使用的标准化格式；而标准化的最终目的是数据实现机器可理解性和可使用性。

目前文本文献是以半结构化形式呈现最新研究数据和信息，同类型领域科研数据分散于大量科技文献和专利文本中，以科研人员可阅读和可理解形式出现，并且由于科研人员书写习惯的不同，文献与文献之间的科研数据表述和组织呈现出差异。因此，需要构建标准化语言和组织格式以规范化文献中的相关数据，满足数据密集型科研范式和科研智能化趋势下科研数据的可获取性、易用性、机器可理解性和重用性。

4 科研智能化趋势下科研数据潜在需求

从上述科研智能化趋势下科研数据处理流程分析过程中发现，AI 模型在

科学研究发展中表现出对多源异构数据的集成、细粒度数据结构化、人机互动数据表示形式的探索、数据关联化挖掘和科研数据类型丰富化的需求。

（1）多源异构数据集成

多源异构数据保证了目标数据及属性获取的全面性，从多角度刻画目标数据知识，有利于 AI 模型特征学习的全面化。从上述案例分析中发现，科研智能化研究数据来源于多个多类型数据库，例如 AlphaFold 模型[47]在利用 PDB 蛋白质结构数据库基础上结合 Uniclust30 蛋白质序列数据以学习蛋白质结构组建的规则。由此，可以看出科研智能化趋势下 AI 模型对多源异构集成数据的需求。多源异构数据的集成不仅有利于 AI 模型的学习，也有利于便利科研人员收集数据，提高科研效率。

（2）细粒度数据结构化

科研智能化趋势下 AI 模型更加注重数据内部隐性规则学习，也即是需要细粒度挖掘数据特征并关联不同类型数据，构建科研人员和机器可理解的结构化数据集，便于科研人员获取和利用。典型案例以文献中的实验方法信息抽取、组织与结构化为代表，如以目标合成物质为核心的实验分解流程图的组织，不仅满足科研人员的细粒度知识学习需求，也满足了 AI 模型对文献数据中隐性知识的学习需求。

（3）人机互动数据表示的探索

现有数据组织形式是以人类可理解形式呈现，满足了科研人员知识学习需求，但科研智能化研究中不仅要注重科研人员的知识学习需求，也需要关注 AI 模型的知识学习需求，需要把现有知识进一步转化为 AI 模型可理解形式，搭建人类语言和机器语言之间的桥梁。以向量化数据库的构建为代表，如 Science Navigator 借助向量计算技术和大语言模型实现了非结构化文献数据的向量化表征、语义搜索、相似度计算，以作为科研智能化研究发展的基础设施支撑。

（4）数据关联化挖掘

科研智能化趋势下的知识发现是以关联识别和挖掘为核心，构建不同层级数据之间的关联关系以实现数据或特征预测的目标，如材料领域材料成分-结构-工艺-性能和化学领域分子结构-性质-功能等复杂构效关系的构建。此外，现有 AI 模型属于“黑箱”模型，在数据关联中的挖掘不具备可解释性，不易于预测结果的理解。因此，面对科研智能化研究需要构建可解释 AI 模型助力关联规则挖掘的可理解性，进一步实现不同领域科研数据隐性关系的显性化。

（5）科研数据类型的丰富化

科研智能化研究中越来越重视实验流程方案的抽取与组织，是科研智能机器人学习的重要数据资源，也是实验组合规律智能分析与发现的核心，其内容是不同实验元素及其数量的组合关系，其提取是对科技文献的提炼和总结，增强了实验方案的机器可读性。现有实验方案的组合是以简单的文本形式组合，检索以单一的实验物质名称为检索核心，不能满足用户以实验目标、实验步骤或实验原理等为核心的检索需求，未来可通过构建领域实验方案知识图谱结合精准推荐技术，根据用户多方面的需求给出推荐内容，辅助科研

人员高效选择实验方案。

5 科研智能化趋势下科研数据服务相关建议

针对文献信息服务机构如何深入参与新科研范式，发挥其在数据服务中的优势，本节基于上述分析为科研数据服务的发展提出以下建议。

（1）加快构建高质量全面化的领域数据网络。数据是科研智能化研究的重要驱动力之一，其获取的便利化、全面性和可用性关乎科研智能化研究的效率和质量。从使用数据来源看，现有科研智能化研究案例中使用的数据多来自于国外数据库、开放数据平台或商业性出版商，国内构建的数据库和开放数据平台使用较少，利用率低，从侧面也说明了国内构建的数据库需要进一步提高数据质量和结构化程度，加强开放性高质量数据库的建立和推广。从使用数据来源数量看，科研智能化研究中使用的数据分散在多个数据平台，需要针对不同平台采取不同的数据获取和分析方法，数据挖掘和分析在科研中的占比较大，降低了科研效率。因此，需要构建统一、高质量、标准化的领域数据平台，集成开放性、商业性和私人性科研数据满足科研智能化对多源数据的需求。

（2）重视文本型数据中实验信息的挖掘。现有科研智能化研究重视文本型数据的细粒度内容挖掘，以构建结构化关联知识。基础科学的智能化研究以实验信息的关联挖掘和结构化为代表，成为自动化流程实验的核心数据。现有实验信息以 Springer • Nature 的 Protocols 实验室指南数据库和 CAS 的 Synthetic Methods 合成试验方法数据库为代表，适用于科研人员查询和学习，但不适用于科研智能化研究的使用和输入，因此，未来需要构建实验流程信息的数据库以支持智能化研究。

（3）关注人机互动语言的探索。AI 模型是科研智能化研究的重要参与者，现有知识时面向人类学习需求服务，其知识内涵和语义关系都需要进一步转化为机器数据表征模式，其特征或表征的模式直接关系知识发现的准确性。因此，需要构建面向不同研究目标的标准化知识表征语言搭建人类知识与机器学习的桥梁。

（4）深化嵌入科研式数据服务模式。在数据密集型时代，我国数据服务模式转变为以数据为核心的数据服务模式，注重前端数据服务中的数据采集、获取和挖掘服务，对科研生命周期中的研究准备服务、数据处理与分析技术选择、数据出版服务关注度较低，导致嵌入式数据服务模式对科研创新的支持和影响力度较低。因此，从科研生命周期出发提供嵌入式数据服务，以实践出发提升科研人员的信息素养和数据素养，激发科研人员的创造力、创新能力和科研能力，才能提升文献信息服务机构在科研智能化趋势中的参与度、影响力和竞争力。

（5）提升图书馆员领域知识和人工智能素养的提升。科研智能化研究是人工智能领域与其他学科领域交叉融合发展的结果，其核心是领域知识与人工智能技术的交融，对图书馆员的数据服务能力也提出了新的要求，在图情领域知识学习的基础上要具备数据分析和挖掘知识成为数据图书馆员，在数据分析和挖掘知识学习基础上要具备领域知识成为学科图书馆员，在领域知

识学习基础上要具备人工智能知识成为智慧化图书馆员，才能提升科研智能化趋势下数据服务的质量。

参考文献：

[1] 都平平, 李雨珂, 陈越. 高校科研数据资产化存储及数据复用权益许可研究 [J]. 图书情报工作, 2022, 66(03): 45-53.

[2] 薛菁华, 徐慧婷, 陈广玉 汇. 全球科研范式数字化转型趋势研究 [J]. 竞争情报, 2022, 18(06): 54-63 %@ 2095-8870 %L 31-2107/G3.

[3] 诸云强, 潘鹏, 石蕾, et al. 科学大数据集成共享进展及面临的挑战 [J]. 中国科技资源导刊, 2017, 49(05): 2-11 %@ 1674-544 %L 11-5649/F %W CNKI.

[4] INITIATIVE M G. Materials Genome Initiative Strategic Plan [Z].

[5] MIKSA T, CARDOSO J, BORBINHA J. Framing the scope of the common data model for machine-actionable data management plans; proceedings of the 2018 IEEE International Conference on Big Data (Big Data), F, 2018 [C]. IEEE.

[6] 陈大庆. 英国科研资助机构的数据管理与共享政策调查及启示 [J]. 图书情报工作, 2013, 57(08): 5-11.

[7] JIANG G, SANTIAGO I A, HANYU G, et al. Automated Chemical Reaction Extraction from Scientific Literature [J]. Journal of chemical information and modeling, 2021.

[8] MEHR S H M, CRAVEN M, LEONOV A I, et al. A universal system for digitization and automatic execution of the chemical synthesis literature [J]. Science, 2020, 370(6512): 101-8.

[9] ADITYA N, CHENRU D, J K H. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal-Organic Frameworks [J]. Journal of the American Chemical Society, 2021.

[10] MYSORE S, KIM E, STRUBELL E, et al. Automatically extracting action graphs from materials science synthesis procedures [J]. arXiv preprint arXiv:171106872, 2017.

[11] OLGA K, HAOYAN H, TANJIN H, et al. Text-mined dataset of inorganic materials synthesis recipes [J]. Scientific data, 2019, 6(1).

[12] COURT C J, COLE J M. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning [J]. npj Computational Materials, 2020, 6(1).

[13] EDWARD K, ZACH J, ALEXANDER V G, et al. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks [J]. Journal of chemical information and modeling, 2020, 60(3).

[14] WANG Z, KONONOVA O, CRUSE K, et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature [J]. Scientific Data, 2022, 9(1): 231.

[15] M J D, E A S, L W E, et al. OSCAR4: a flexible architecture for chemical text-mining [J]. Journal of cheminformatics, 2011, 3(1).

[16] LEZAN H, M J D, NICO A, et al. ChemicalTagger: A tool for semantic text-mining in chemistry [J]. Journal of Cheminformatics, 2011, 3(1).

[17] C S M, M C J. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature [J]. Journal of chemical information and modeling, 2016, 56(10).

[18] XINTONG Z, STEVEN L, SEMION S, et al. Text to Insight: Accelerating Organic Materials Knowledge Extraction via Deep Learning [J]. Proceedings of the Association for Information Science and Technology, 2021, 58(1).

[19] PANG N, QIAN L, LYU W, et al. Using pretraining and text mining methods to automatically extract the chemical scientific data [J]. Data Technologies and Applications, 2021, 56(2).
[20] PARK H, KANG Y, CHOE W, et al. Mining Insights on Metal‑Organic Framework Synthesis from Scientific Literature Texts [J]. Journal of Chemical Information and Modeling, 2022, 62(5): 1190-8 %@ 549-9596.
[21] ZACH J, EDWARD K, SOONHYOUNG K, et al. A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction [J]. ACS central science, 2019, 5(5).
[22] BURGER B, MAFFETTONE P M, GUSEV V V, et al. A mobile robotic chemist [J]. Nature, 2020, 583(7815).
[23] W C C, A T D, M L J A, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning [J]. Science (New York, NY), 2019, 365(6453).
[24] M G J, LIVA D, VINCENZA D, et al. Controlling an organic synthesis robot with machine learning to search for new reactivity [J]. Nature, 2018, 559(7714).
[25] ALESSANDRA T, PHILIPPE S, ANTONIO C, et al. Unassisted noise reduction of chemical reaction datasets [J]. Nature Machine Intelligence, 2021, 3(6).
[26] CHRISTOPH S, CHRISTIAN H, STEFAN K, et al. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics [J]. Current pharmaceutical design, 2006, 12(17).
[27] SCHWALLER P, HOOVER B, REYMOND J-L, et al. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions [J]. Science Advances, 2021, 7(15): eabe4166 %@ 2375-548.
[28] 施思齐, 涂章伟, 邹欣欣, et al. 数据驱动的机器学习在电化学储能材料研究中的应用 [J]. 储能科学与技术, 2022, 11(03): 739-59.
[29] FALAK N, ANIRUDDH H, JANAMEJAYA C, et al. A generative adversarial network‑based synthetic data augmentation technique for battery condition evaluation [J]. International Journal of Energy Research, 2021, 45(13).
[30] KIM E, HUANG K, JEGELKA S, et al. Virtual screening of inorganic materials synthesis parameters with deep learning [J]. npj Computational Materials, 2017, 3(1).
[31] LOOKMAN T, BALACHANDRAN P V, XUE D, et al. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design [J]. npj Computational Materials, 2019, 5(1).
[32] VISHU G, KAMAL C, FRANCESCA T, et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data [J]. Nature Communications, 2021, 12(1).
[33] CHERRINGTON M, THABTAH F, LU J, et al. Feature selection: filter methods performance challenges; proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), F, 2019 [C]. IEEE.
[34] BENNASAR M, HICKS Y, SETCHI R. Feature selection using Joint Mutual Information Maximisation [J]. Expert Systems With Applications, 2015, 42(22).
[35] KHAIRE U M, DHANALAKSHMI R. Stability of feature selection algorithm: A review [J]. Journal of King Saud University - Computer and Information Sciences, 2019, (prepublish).
[36] WEIHAO Z, TEHILA E, TINGTING W, et al. Multi-feature based network revealing the structural abnormalities in autism spectrum disorder [J]. IEEE Transactions on Affective Computing, 2019.
[37] JIHENG F, MING X, XINGQUN H, et al. Machine learning accelerates the materials discovery [J]. Materials Today Communications, 2022, 33.

[38] 刘悦, 马舒畅, 杨正伟, et al. 面向材料领域机器学习的数据质量治理 [J]. 硅酸盐学报, 2023, 51(02): 427-37.
[39] COLEY C W, JIN W, ROGERS L, et al. A graph-convolutional neural network model for the prediction of chemical reactivity [J]. Chemical science, 2019, 10(2): 370-7.
[40] KANGJIE L, YOUJUN X, JIANFENG P, et al. Automatic retrosynthetic route planning using template-free models [J]. Chemical Science, 2020, 11(12).
[41] OPENAI. GPT-4 Technical Report [J]. ArXiv, 2023, abs/2303.08774.
[42] CHARLIN L, ZEMEL R. The Toronto paper matching system: an automated paper-reviewer assignment system [J]. 2013.
[43] YUAN W, LIU P, NEUBIG G. Can we automate scientific reviewing? [J]. Journal of Artificial Intelligence Research, 2022, 75: 171-212.
[44] 沈志宏, 张晓林, 郑晓欢 中, et al. PARIS 原则：开放协作环境下科学数据的可用性 [J]. 大数据: 1-16 %@ 2096-0271 %L 10-1321/G2 %U
https://kns.cnki.net/kcms/detail/10..g2.20230111.1741.001.html %W CNKI.
[45] ZHU Q, ZHANG F, HUANG Y, et al. An all-round AI-Chemist with a scientific mind [J]. National science review, 2022, 9(10).
[46] 吴思远, 王宇琦, 肖睿娟, et al. 电池材料数据库的发展与应用 [J]. 物理学报, 2020, 69(22): 9-16.
[47] JOHN J, RICHARD E, ALEXANDER P, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873).

(通讯作者: 韩涛　E-mail:hant@mail.las.ac.cn)

[作者贡献声明]

张婧睿：文献调研与整理；撰写论文；
韩涛，孙蒙鸽：修订论文，审核论文。